Genomics **67,** 201–209 (2000)
doi:10.1006/geno.2000.6241, available online at http://www.idealibrary.com on IDEAL®

Multivariate Measurement of Gene Expression Relationships

Seungchan Kim,* Edward R. Dougherty,* Yidong Chen,† Krishnamoorthy Sivakumar,‡
Paul Meltzer,† Jeffery M. Trent,† and Michael Bittner†,[1]

*Department of Electrical Engineering, Texas A&M University, College Station, Texas 77843; †Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892; and ‡Department of Electrical Engineering, Washington State University, Pullman, Washington 99164

Received November 15, 1999; accepted May 2, 2000

The operational activities of cells are based on an awareness of their current state, coupled to a programmed response to internal and external cues in a context-dependent manner. One key goal of functional genomics is to develop analytical methods for delineating the ways in which the individual actions of genes are integrated into our understanding of the increasingly complex systems of organelle, cell, organ, and organism. This paper describes a novel approach to assess the codetermination of gene transcriptional states based upon statistical evaluation of reliably informative subsets of data derived from large-scale simultaneous gene expression measurements with cDNA microarrays. The method finds associations between the expression patterns of individual genes by determining whether knowledge of the transcriptional levels of a small gene set can be used to predict the associated transcriptional state of another gene. To test this approach for identification of the relevant contextual elements of cellular response, we have modeled our approach using data from known gene response pathways including ionizing radiation and downstream targets of inactivating gene mutations. This approach strongly suggests that evaluation of the transcriptional status of a given gene(s) can be combined with data from global expression analyses to predict the expression level of another gene. With data sets of the size currently available, this approach should be useful in finding sets of genes that participate in particular biological processes. As larger data sets and more computing power become available, the method can be extended to validating and ultimately identifying biologic (transcriptional) pathways based upon large-scale gene expression analysis.

INTRODUCTION

One of the fundamental goals of genomics research is to understand the ways in which the networks of gene activity are integrated in the cell. This is currently a very difficult problem, if for no other reason than the sheer number of possible interactions. Traditional biochemical and genetic characterizations of genes do not allow rapid sifting of these possibilities to identify the genes involved in processes or the control mechanisms employed. On the other hand, when methods exist to focus genetic and biochemical characterization methods on a smaller number of genes likely to be involved in a process, progress in finding the relevant interactions and controls can be substantial.

Our earliest understandings of the mechanics of cellular gene control were derived in large measure from studies of just such a case, metabolism in simple cells. In metabolism, it was possible to use biochemistry to identify stepwise modifications of the metabolic intermediates and genetic complementation tests to identify the genes responsible for catalysis of these steps and control of their expression. Standard methods of characterization guided by some knowledge of the connections could thus be used to identify process components and controls. Starting from the basic outline of the process, molecular biologists and biochemists were able to build up a very detailed view of the processes and regulatory interactions operating within the metabolic domain.

In contrast, for most cellular processes, general methods to implicate likely participants and to suggest control relationships have not emerged. The resulting inability to produce overall schemata for most cellular processes has meant that gene function is, for the most part, determined in a piecemeal fashion. Once a gene is suspected of involvement in a particular process, research focuses on the role of that gene in a very narrow context. This typically results in the full breadth of important roles for well-known, highly characterized genes being slowly discovered. A particularly good example of this is the relatively recent appreciation that oncogenes such as Myc can stimulate apoptosis in addition to proliferation (Evan and Littlewood, 1998).

The recognition of this bottleneck has stimulated the field's appetite for methods that can provide a wider

[1] To whom correspondence should be addressed. Telephone: (301) 496-7989. Fax: (301) 402-3241. E-mail: mbittner@nhgri.nih.gov.

0888-7543/00 $35.00

experimental perspective on how genes interact. Recently, methods for carrying out large-scale surveys of gene transcripts using cDNA microarrays that can produce enormous data sets concerning transcriptional levels have been described (Schena *et al.*, 1995, 1996; DeRisi *et al.*, 1996, 1997; Wodicka *et al.*, 1997). As these measurements are a snapshot of the types and levels of transcripts required to achieve or maintain the cell state being observed, they are a *de facto* source of information about transcript interactions involved in gene regulation.

Analysis of these data can take two routes, univariate, gene by gene analysis, or multivariate, analysis of interactions among many genes simultaneously. To date, the bulk of published analysis of expression profiles has involved the very useful, but computationally basic, analysis of data sets based entirely on the univariate correlation of gene expression changes (Eisen *et al.*, 1998; Spellman *et al.*, 1998; Tavazoie *et al.*, 1999). Correlation can identify common elements of a cell's response to a particular stimulus and thus discern some groups of genes. However, correlation does not address the fundamental problem of determining the sets of genes whose actions and interactions drive the cell's decision to set the transcriptional level of a particular gene. Because transcriptional control is accomplished by a complex method that interprets a variety of inputs (McAdams and Shapiro, 1995; Evan and Littlewood, 1998; Yuh *et al.*, 1998), the development of analytical tools that detect multivariate influences on decision-making present in complex genetic networks is essential. To carry out such an analysis, one needs both an analytical method and sufficient data.

One plausible way of gaining insight into multivariate influences on a gene in a cell is to study its expression pattern in a variety of defined circumstances and cellular contexts, as it responds to its environment and to the actions of other genes. Engineers have developed very powerful mathematical approaches to modeling the relationships that govern the function of a component of a complex system by inferences derived from comparisons of its state to the states of other system components during the operation of the system. This process is referred to as "reverse engineering." Considerable effort has been expended in evaluating how applicable this approach is to a biological network (reviewed in Szallasi (1999)). These methods are not easily applicable to biology. The first steps in reverse engineering are to model the system in some basic fashion and then to obtain a very large number of precise measurements of the behavior of the components as the system operates. These requirements are quite constricting even when a very simple model is to be evaluated. Observation of 100 state transition pairs were estimated to be required to model a Boolean network containing 50 elements with no more than 3 inputs per element (Liang *et al.*, 1998). Such data sets could conceivably be generated in the case where one was studying an *in vitro* system, say a differentiating cell line. In this case, the starting point is a homogenous set of cells, where temporal sampling along the course of development could easily be carried out. It would also be necessary that it already be sufficiently clear which genes played a part in the process to be studied, so that the relevant components were sure to be measured. Clearly, this kind of modeling could not be carried out in the study of the progression of a human disease, where the samples would be inhomogeneous, starting from somewhat different initiation states, and not in temporal synchrony (Akutsu *et al.*, 1999).

A further complication exists in the precision of the measurement required for such modeling. The detection technique would need to be able to detect differences in the level of the components over most of the measurements reliably, so that the shared information in the system could be easily evaluated. Expression profiling is an immature analytical tool, the results it will provide in the next few years will not be precise to within a few percent. This inexactitude combined with the expense of making the determinations and the problems of obtaining or preparing samples that will sample the widest possible number of cell states strongly limits the feasible multivariate approaches that can be applied to transcript profile data.

What can be done when the data available are imprecise, when not all the genes involved in a given process may be measured in a given experimental series, and when the set of genes involved in the process being studied is not fully known? To obtain the best yield from such data, we propose a modest multivariate method aimed at finding the most manifest network relationships while exploiting the most extensive use of prior knowledge possible. The approach is to find associations between gene expression patterns by measuring the determination (predictive relation) between the transcriptional levels of a small gene set and the transcriptional state of another gene. One of the simplifications that makes this approach practical with partial data is that it does not require a mechanistic model of the way in which the genes influence one another as a basis of evaluation. By accepting this degree of indeterminacy, it is possible to obtain an estimate of the extent to which their states are bound, sensitive even when only some of the sources of influence are present in the observed data set. The trade-off is that the method provides no exact information on the order or proximity of the connections; one only obtains a sense of whether genes are more or less interactive.

This method is geared toward application to data culled for the most statistically reliable observations, from experiments designed to sample changes due to particular processes. The goal is not to produce a full and quantitatively accurate model of the working of the network, but to speed the processes of identifying unexpected new components of already identified processes and of finding unexpected links between processes not previously known to be coordinated. A major improvement over previous tools is the ability to incorporate knowledge of other conditions (such as the application of particular stimuli or the presence of inactivating gene mutations) as predictive elements affecting the of expression level of a given gene. The method is designed to help focus the powerful traditional characterization methods on small sets of genes involved in particular processes, so that such efforts are maximally productive.

MATERIALS AND METHODS

Preparing data for codetermination analysis. As a first step in carrying out nonlinear genomic prediction on gene expression profiles, data complexity is reduced by thresholding the changes in transcript level into ternary expression data: [-1 (down-regulated), $+1$ (up-regulated), or 0 (invariant)]. This simplification allows us to ensure a high and uniform level of certainty in specifying which genes have undergone significant changes in levels of message expression across large numbers of individual microarray experiments. The use of very simple, ternary data representations also makes it possible to use data from differing assay platforms, such as Northern or dot blot determinations, or to include entirely different forms of data, such as application of a stimulus or the functional status of a gene, when these types of data are available.

To find connections between genes, enough differing conditions must be sampled that the independent functioning of different genetic networks can be detected. This amount of sampling requires data from numerous individual experiments. In the example of cDNA microarrays, when looking across numbers of arrays, the absolute intensity of signal detected by each element of the detector in this hybridization-based assay varies based on both the EST printing process and the processes of preparing and labeling the cDNA pools. This problem has been solved by recourse to internal standardization. By using fluorescence detection schemes, a test and a reference probe can be simultaneously hybridized to each array, and expression levels can be reported as a ratio relative to the reference probe. Further reliability can be obtained by applying the same reference probe throughout an array study and by using the variance of a large set of “housekeeping” genes to estimate the statistical significance of observed ratios. An algorithm that first calibrates the data internally to each microarray and statistically determines whether the data justify the conclusion that expression is up-regulated or down-regulated with 99% confidence is currently in use to detect significant changes in transcript levels (Chen *et al.*, 1997). In the case of expression data from quantitative dot blot analysis, up-regulation and down-regulation were judged to be significant at a twofold change in expression from the reference sample (Amundson *et al.*, 1999). Requiring a high confidence level ensures that the logical values -1 and 1 represent significant down- and up-regulation, respectively, and that they are very unlikely to result from natural variability within expression levels or from experimental variability. We are especially concerned with avoiding false conclusions that a certain vector of expression levels within a gene set predicts up- or down-regulation.

Finding predictive relationships. The basic tool for studying shared state determination is one that makes it possible to explore systematically whether orderly relationships exist between genes. The explicit question to be answered is whether a set of arithmetic or logical rules can constructed that allow one to predict the state of one variable, the predictive target, based on the known state of another set of variables, the set of predictors, with some degree of accuracy. One set of tools that mathematicians have developed to explore this type of question is termed perceptrons (Rosenblatt, 1962; Bishop, 1995; Astola and Kuosmanen, 1999). Perceptrons form a class of nonlinear operators that share some properties of linear predictors. We use them here because of their simplicity and the relatively small amount of data required to design them. A schematic figure of one type of perceptron, a linear predictor, that can be used to search for state predictive rules among genes, whose relative abundances are known over a series of samples, is shown in Fig. 1.

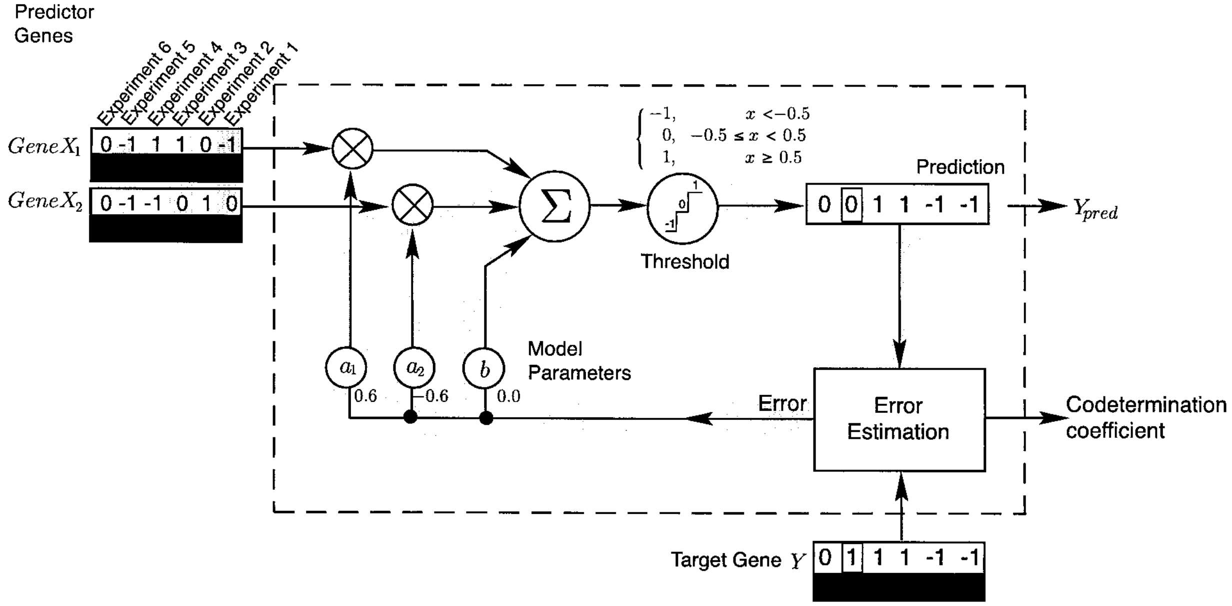
The process depicted in Fig. 1 is the operation of a recursive algorithm, which can search through the possible settings of a linear predictor of the general form

$$Y_{\text{pred}} = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b,$$

to find settings that produce the greatest accuracy of prediction. The algorithm shown attempts to find a simple, mathematical expression that can use the expression data obtained for two genes (X_1, X_2) over a series of experiments to predict the behavior observed for a third gene (Y) in the same series of experiments. For this form of perceptron, the process starts with an estimation of a weighting factor (a_i) to be applied to the expression value for each gene (X_i), an estimation of an overall offset factor (b) to be added to the weighted sums, and estimation of threshold values that will determine the value of the prediction given the final sum. After initial estimations are supplied, the algorithm is run, and the predicted values for Y based on the observed values for the genes are compared to the observed values for Y. Based on the extent and direction of the prediction errors, the estimations are revised, and the algorithm is run again. The attempt at incremental improvement continues for a fixed number of cycles. The accuracy after adjustment is compared with the accuracy when using the initial estimates, and then the best prediction is reported. An example set of gene states, weights, offset, and thresholds is provided in Fig. 1. A table showing the predictive results of applying these settings in the given perceptron and the accuracy of the predictions is contained within Fig. 1. The mathematical details of producing the mathematical and logical perceptrons used in this study are as described in Kim *et al.* (in press).

Measuring the accuracy of expression level predictions. The measurement of fit in a multivariate, biological setting needs to accommodate situations in which predictive strengths vary from poor to good, and only partial predictive power is provided by individual components of the predictor. This requirement can be satisfied by moving from estimating predictive error based on the mean-square error (MSE) to measurement of the normalized mean-square error (NMSE). If we consider only the linear correlation ρ between X and Y, then our understanding concerns the prediction of Y from X via a linear formula, $Y_{\text{pred}} = aX + b$, where Y_{pred} estimates Y via X. Based on the MSE, which is the expected squared difference between Y_{pred} and Y, if X and Y are jointly normally distributed, then the error of the best linear predictor is $\sigma_Y^2(1 - \rho^2)$, with σ_Y^2 being the variance of the target Y. The NMSE is $1 - \rho^2$ and is obtained by dividing the MSE by σ_Y^2. If $|\rho| \approx 1$, then there is very small normalized error. Normalization is thus important in allowing us to observe smaller increments of prediction. This is extremely useful in the biological case where context-based rules will be expected to provide varying outputs from the same inputs, due to differences in the state of the system when it receives the inputs. This “particularization” of responses will have the effect of producing a very large set of “rules,” any of which might be operating at a low frequency in the population of cells sampled. To recognize predictable behavior in such systems, one needs a very sensitive analytic tool, since even if the MSE is small, the degree to which knowledge of X affects our knowledge of Y need not be great.

The use of NMSE allows very sensitive detection of the ability of a predictor to increase, even partially, the accuracy of prediction of the target. Using no predictor variables, the best MSE predictor of the target Y is μ_Y, the mean of Y, and the MSE of prediction of Y by μ_Y is σ_Y^2. The gain in prediction by using X is $\sigma_Y^2 - \sigma_Y^2(1 - \rho^2) = \sigma_Y^2\rho^2$. Normalization of the gain by σ_Y^2 yields ρ^2, which in this context is called the coefficient of determination. The “determination” terminology does not indicate the physical means of determination of Y by X; rather, it refers to the reduced variation in $Y - Y_{\text{pred}}$, as opposed to the variation of Y. The closer ρ^2 is to 1, the smaller is the variance of $Y - Y_{\text{pred}}$, and the more $Y - Y_{\text{pred}}$ is determined. Y_{pred} approximates the random behavior of Y, not just its centrality. Note that predicting



Predictor		*Prediction*		*Square Error*
X_1	X_2	Y_{pred}	Y	$e^2 = (Y_{pred} - Y)^2$
−1	−1	0	1	1
−1	0	−1	−1	0
−1	1	−1	−1	0
0	−1	1	1	0
0	0	0	0	0
0	1	−1	−1	0
1	−1	1	1	0
1	0	1	1	0
1	1	0	0	0

FIG. 1. Schematic view of the operation of a perceptron. The perceptron is shown at the end of a cycle of attempting to improve the prediction of the state of gene Y, using information on the states of genes X_1 and X_2. The data set consists of message abundance readings for all three genes in a series of six experiments. The expression values have been thresholded to ternary values, −1 (down-regulated relative to a fixed reference), +1 (up-regulated), or 0 (invariant). Weights a_1 and a_2 are applied to the values of X_1 and X_2, and an offset value is applied at b. Adjusted thresholding rules to interpret the value of the sum as a predicted value for gene Y are shown above the threshold symbol. The boxed values in the predicted value of Y and the observed value of Y show the error in prediction made with these settings. The complete table of predictions that would be made by the perceptron with these parameters is shown beneath the schematic. Further cycles of parameter adjustment would be undertaken to see if an entirely correct rule could be produced.

Y by its mean might yield small MSE, but the variance of $Y - \mu_Y$ is the same as the variance of Y. Thus, predictions measured in terms of the NMSE allow us to discern partial predictive value confidently, so long as it is significant relative to the variance of Y.

RESULTS

Representations of Predictive Results

Results will be presented as arrow plots, with the target gene at the right and the chained predictors plotted to the left. The determination achieved by adjoining a predictor gene is placed on the arrow following it. For instance, in Fig. 2, Predictor 1 achieves determination θ_1 for predicting the target gene, using Predictors 1 and 2 together achieves determination θ_2, and using Predictors 1, 2, and 3 together achieves determination θ_3.

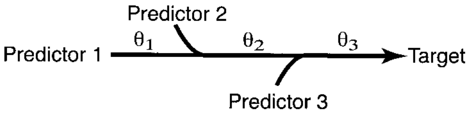


FIG. 2. Determination diagram (template).

Testing the Perceptron with Control Data

As a blind control, expression patterns for two fictitious genes were created. Rules were made for each gene dependent on other gene states in the set, and different levels of noise were introduced in the two controls. The noise is added to make the problem of prediction more realistic, as it is expected that few if any multigene systems will exhibit gene states perfectly predictable by a very limited number of genes. The first, AHA, had a pattern dependent on the rule set: up-regulated if p53 is functional, down-regulated if RCH1 and p53 are deficient. In the absence of noise, the rule would produce 15 instances of up-regulation and 5 instances of down-regulation. The data set generated for this gene included 13 of the 15 up-regulations and 5 of the 5 down-regulations, a low-noise test. The rule set devised for the second gene, OHO, was as follows: up-regulated if MDM2 is up-regulated and RCH1 is down-regulated, and down-regulated if p53 is down-regulated and REL-B is up-regulated. In the absence of noise, this set of rules would produce 4 instances of up-regulation and 5 instances of down-reg-

TABLE 1

Ternary Expression Data for IR Responsive Genes and Synthetic Control Genes

		Gene														Condition		
Cell line	Condition	RCH1	BCL3	FRA1	REL-B	ATF3	IAP-1	PC-1	MBP-1	SSAT	MDM2	p21	p53	AHA	OHO	IR	MMS	UV
p53 proficient																		
ML-1	IR	−1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
ML-1	MMS	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0
Molt4	IR	−1	0	0	1	1	0	1	0	0	1	1	1	1	1	1	0	0
Molt4	MMS	0	0	1	0	1	0	0	0	0	0	1	1	1	0	0	1	0
SR	IR	−1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0
SR	MMS	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0
A549	IR	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0
A549	MMS	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	1	0
A549	UV	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1
MCF7	IR	−1	0	1	1	0	0	0	0	0	1	1	1	0	1	1	0	0
MCF7	MMS	0	0	1	0	1	0	0	0	0	1	1	1	1	0	0	1	0
MCF7	UV	0	0	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1
RKO	IR	0	1	0	1	1	1	1	0	0	1	1	1	1	0	1	0	0
RKO	MMS	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	1	0
RKO	UV	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1
p53 deficient																		
CCRF-CEM	IR	−1	1	1	1	1	0	1	0	0	0	0	−1	−1	0	1	0	0
CCRF-CEM	MMS	0	0	0	0	1	0	0	0	0	0	0	−1	0	0	0	1	0
HL60	IR	−1	1	0	1	1	0	1	0	1	0	1	−1	−1	−1	1	0	0
HL60	MMS	0	0	1	0	1	0	0	0	0	1	1	−1	0	1	0	1	0
K562	IR	0	0	0	0	0	0	0	0	0	0	0	−1	0	0	1	0	0
K562	MMS	0	0	0	0	1	0	0	0	0	0	0	−1	0	0	0	1	0
H1299	IR	0	0	0	1	0	0	1	0	0	0	0	−1	0	0	1	0	0
H1299	MMS	0	0	0	0	1	0	0	0	0	0	1	−1	0	1	0	1	0
H1299	UV	0	0	0	0	1	0	1	0	0	0	1	−1	0	1	0	0	1
RKO-E6	IR	−1	1	0	1	0	1	1	0	0	0	0	−1	−1	0	1	0	0
RKO-E6	MMS	−1	0	0	0	1	0	0	0	0	0	1	−1	−1	1	0	1	0
RKO-E6	UV	−1	0	0	0	1	0	0	0	0	0	1	−1	−1	1	0	0	1
T47D	IR	0	0	0	1	0	0	0	0	0	0	1	−1	0	−1	1	0	0
T47D	MMS	0	0	0	0	1	0	0	0	0	0	1	−1	0	1	0	1	0
T47D	UV	0	0	0	0	1	0	0	0	0	0	1	−1	0	1	0	0	1

ulation. The data set generated for OHO had the 4 expected up-regulations plus 7 unpredicted up-regulations and only 2 of the 5 predicted down-regulations, a very noisy data set. The ternary data of the survey and controls are given in Table 1, where the conditions IR, MMS, and UV have the values 1 or 0, depending on whether the condition is in effect or not in effect, respectively.

For the control gene AHA, the perceptron identified the p53 and RCH1 components of the transcription rule set (up-regulated if p53 is functional, down-regulated if RCH1 and p53 are deficient). Substantial gains in accuracy of prediction were achieved by inclusion of these two genes in the prediction (Fig. 3A). The addition of PC1 is probably increasing accuracy through a mechanism correcting for the introduced noise.

Since many violations of the rules (up-regulated if MDM2 is up-regulated and RCH1 is down-regulated, and down-regulated if p53 is down-regulated and REL-B is up-regulated) were introduced into the data set for the OHO gene, it was expected that only limited determination by these genes would be observed. This expectation is met, only very marginal gains are associated with MDM2 and RCH1, and no significance is found for combinations involving p53 and REL-B (Fig. 3B). The controls clearly show that the perceptron method is capable of detecting a degree of codetermination even when the rules are not perfectly followed and that the codetermination coefficients provide useful metrics for visualizing the extent to which rules are followed.

Applying the Perceptron to Blot and Microarray Gene Expression Data

Tests of the ability of the perceptron to detect associations based on changes in transcription level have been performed in the context of responsiveness to genotoxic stresses. As a result of a microarray study surveying transcription of 1238 genes during the re-

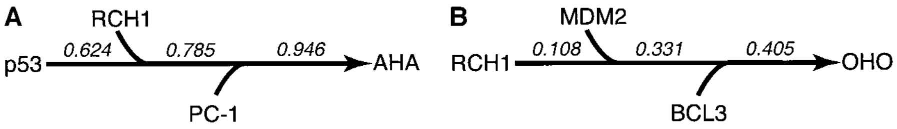


FIG. 3. Determination diagrams of best predictors for artificial, synthetic control, genes.

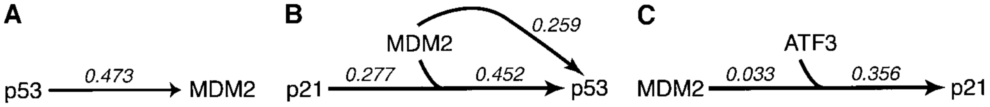


FIG. 4. Determination diagrams of best predictors where there is consistency with biological information.

sponse of a myeloid line to ionizing radiation (Amundson *et al.*, 1999), 30 genes not previously known to participate in response to IR were found to be responsive. To characterize further the responsiveness of these genes to genotoxic stresses, the responsiveness of a subset of 9 of them was examined by blot assays in 12 cell lines stimulated with ionizing radiation, a chemical mutagen (methyl methane sulfonate, MMS), or ultraviolet radiation. The cell lines were chosen so that a sampling of both p53-proficient and p53-deficient cells would be assayed. The data set thus consists of measurements of transcript abundance for 12 genes under 30 conditions. Abundance for one of the conditions was measured with microarrays, and the remaining 29 were measured by dot blot.

It can immediately be seen that the genes included in this survey are not uniformly regulated in the various cell types. All genes showed an up- or a down-regulation in at least one other cell type; however, the extent of changes registered across the lines were quite variant. Such a varied response reflects the distinct ways in which different cells respond to the same external stimuli based on their own internal state and is therefore a useful test set for this methodology.

One technical aspect that emerges in studying these cases is the clear relationship between the number of changes observed for a particular gene in the data sets and the quality of prediction that can be achieved. Since the perceptron operates by finding rules that can relate changes observed in one gene with changes in others, it is necessary that the gene being predicted change a significant number of times in the set of observations to obtain a meaningful prediction. A determination of the level of confidence that can be assigned for a prediction of a gene exhibiting a particular fraction of change will need to be developed. In the current case, we will arbitrarily limit the set of genes for which a prediction will be made to those exhibiting at least 4 changes in the set of 30 observations, eliminating MBP1 and SSAT as targets of prediction.

Predictions Involving Genes with Known Relationships

In those cases where existing biological information provided expectations about the predictive relationships between genes in the test set, the perceptron predictions conformed to these expectations. The biological expectation would be that MDM2 would be incompletely predicted by p53. This expectation is met (Fig. 4A). Additions of further genes to p53 do not increase the accuracy of the prediction. Similarly, as it is known that p53 is influential, but not determinative of the up-regulation of both p21 and MDM2, some level of prediction of p53 should be possible by a combination of these two genes. This expectation is also met (Fig. 4B). Moreover, as p21 shows both p53-dependent and p53-independent regulation in response to genomic damage (Gorospe *et al.*, 1996), it was expected that the p53 component would not be recognized by the algorithm. p53 was not selected for the predictor. The algorithm chose the somewhat similar pattern of expression exhibited by ATF3, with some supplementary information from the MDM2 pattern as the best predictor of p21 (Fig. 4C). The prediction carries borderline significance.

Predictions Involving Genes where Interrelatedness Is Not Established

In surveying the newly found, IR responsive genes FRA1, ATF3, REL-B, RCH1, PC1, IAP-1 and MBP-1, we see two very distinct patterns of interrelationship. For the genes FRA1 and ATF3, all predictions are weak, having determination coefficients less than 0.1. FRA1 (Fos-related antigen) is an immediate-early gene, induced by serum stimulation (Cohen and Curran, 1988). It is quite possible that this gene is an early component of a more generic stress-induced genetic network that is largely independent of the radiation responses catalogued to date. The small number of genes considered in this experiment makes it likely that some will exhibit no relation to others in the experiment. Similar considerations may apply to ATF3, activating transcription factor 3. ATF3 has been shown to be widely inducible by a variety of stresses, including wounding, phorbol ester stimulus, $CHCl_4$ and alcohol exposure, ischemia/reperfusion, and brain seizure (Chen *et al.*, 1994, 1996). It is clearly far more responsive to MMS and UV induction than to IR. This gene might be an early responder whose downstream targets are not represented in this survey.

The other set of relationships seen among the newly found IR responsive genes involves predictions that seem to link the behavior of REL-B, RCH1, PC-1, MBP-1 BCL3, and IAP-1. The perceptron finds a variety of shared expression behaviors within this set. A sampling of determination diagrams showing mutual predictability among these genes is shown in Fig. 5. In examining the full list of predictions involving these genes, it became clear that in cases where an inducing condition was included in the prediction, the extent of predictability was higher when the condition was exposure to ionizing radiation. As shown in Fig. 6, significant determination could be observed even when ionizing radiation itself was not the most potent predictor.

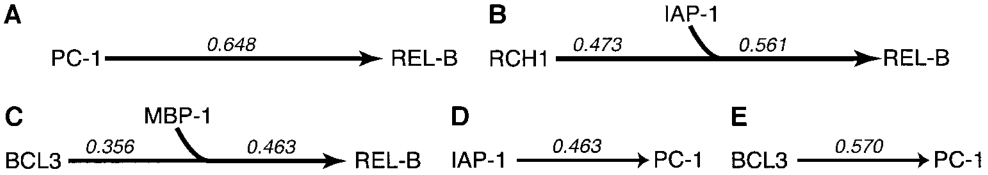


FIG. 5. Determination diagrams showing mutual predictability among some genes newly found to be IR responsive.

When this list of genes is studied with an eye to IR responsiveness, it becomes apparent that they share an overall trend to show expression level changes in response to ionizing radiation rather than to UV or MMS. The extent of this preference, and the degree to which it differs from the more generally responsive behavior shown by FRA1 and ATF3, is detailed in Table 2.

These results provide an example of the ability of this form of analysis to highlight subtleties in the data. Even though MBP1 and SSAT responded to IR only at the very low rate of 17% of the possible chances, they responded exclusively to this stimulus, and not to the other genotoxic stimuli, and were thus associated with other genes showing a similarly high preference to respond to IR. Among this cohort of preferential IR responders, the general pattern of the most predictive sets is a large first-step gain in determination with little further gain by inclusion of other genes. This pattern is consistent with a very loose degree of connection, primarily reflecting the common response pattern. Perhaps all the responses derive from distant branchings of some earlier network signal, or perhaps a number of networks are simultaneously engaged at an early time by exposure to ionizing radiation. This is consistent with the information available for the function of these genes, which indicates that they are very unlikely to constitute a simple functional pathway. REL-B is known to be a modifier of NFκB, a transcription activator commonly induced in response to genotoxic shock (Liou *et al.*, 1994; Ivanov *et al.*, 1995). RCH1 (recombination activating gene cohort 1) acts as a promoter of docking of cytosolic substrates that have nuclear localization signals (Gorlich *et al.*, 1995). PC1 (prohormone convertase 1) is a subtilisin-like proprotein processing enzyme, found to be frequently highly expressed in carcinoid tumors (Creemers *et al.*, 1992).

DISCUSSION

From metabolism to cell cycle control, the organism/cell is involved in the constant monitoring and passage of information between its various components. The attempt to understand how this extraordinary level of interconnection and integration of cellular activities functions in healthy cells and fails in diseased cells poses questions of enormous complexity. Our currently available analytic methods to examine gene networks function primarily to extend of our knowledge of recognized interconnections among characterized genes, providing the gradual addition of new genes to established networks. However, this approach typically suppresses to the absolute minimum the impact of other cellular processes on the specific pathway being studied, so that the effects of alteration of one or a few genes may be unambiguously observed. A clear sense of the way in which gene interactions become less and less readily interpretable when their responses under different conditions are viewed is provided in Fig. 7. Here the predictability of p53, MDM2, and p21 can be seen to decline steeply as one's view broadens from the way in which they respond in a single stress response to their behavior in multiple stress responses. As a result, these approaches are an excellent way to derive detailed information about established relationships, but are poor at discovering new relationships.

In contrast, a unique opportunity now exists at the intersection of sample EST sequencing, an undirected gene discovery method, and microarray analysis of gene expression, an undirected function discovery method. The data developed with these methods are suitable substrates for the development of analytic methods focused on discovering relationships between genes. The method described in this paper allows the modeling of transcription information against the functional integration of gene activity resulting from defined cellular stimuli (ionizing radiation or gene mutation status). The analysis attempts to bring together two central observations about cells as systems. First, as cells modify themselves to respond to their circumstances, portions of their reconfigurations involve concerted changes in the levels of mRNA for genes involved in the response. Second, the collaborative efforts

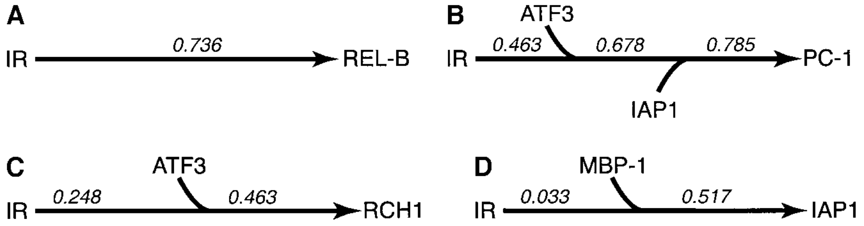


FIG. 6. Determination diagrams for IR responsive genes when IR is included as a predictor variable.

TABLE 2

Preferential IR Responsiveness of a Subset of Genes to Ionizing Radiation

Gene	Observed changes	IR-induced changes	% IR-induced changes	% Possible IR responses
RELB	11	10	91	83
PC-1	9	8	89	67
RCH1	9	7	78	58
BCL3	5	5	100	42
IAP-1	4	4	100	33
MBP1	2	2	100	17
SSAT	2	2	100	17
FRA1	7	3	43	12
ATF3	24	6	25	50

of genes span a large range of connectivity, from the obligate, such as partners in a particular catalytic complex, to the highly contingent, such as effectors of DNA repair, which are mobilized when damage occurs on the basis of the specific type of damage sustained. It is therefore expected that changes in mRNA levels will, for some sets of genes, reflect their level of functional coupling. Further, if the relative abundance of the messages for these genes is observed over a wide sampling of cell states, then it will be possible to rank the tightness of coupling between these genes. This can be accomplished by determining how accurately the states of a set of genes predict the state of some other gene. The higher the degree of relationship, i.e., the more codetermined the set of genes is, the more accurate the prediction.

A mathematical method for statistically assessing codetermination has been described. It has been demonstrated that a nonlinear perceptron can be used to ferret out known and constructed relationships, to provide useful measures of the strengths of codetermination, and to disclose subtle similarities of transcriptional activity. The method has the flexibility to allow predictions to be formulated and evaluated based not only on expression data, but also on the conditional functionality of genes and on applied external stimuli. The combination of insensitivity to codetermining mechanism, capacity for multicomponent prediction, mixing gene states, and other influences, and capability of detecting imperfect codetermination make this methodology well suited to searching for connections between genes with the kinds of data sets currently being generated with array-based technology. However, this type of statistical approach is not restricted to array data, but could be equally well applied to other components, such as protein levels, where the presence and abundance of the component can be confidently measured.

To apply this approach broadly, a number of developments are critical. The amount of computation required to systematically produce codetermination estimates for the large numbers of genes that show altered abundance in a series of array experiments is quite massive, but approachable. To reduce the number of spurious or chance predictions, the number of genes used as targets and predictors should be stringently filtered to ensure that the changes observed are significant and that a sufficient number of changes are observed in the series to allow for strong prediction. Such filtration would typically reduce the pool of genes suitable for evaluation to less than 10% of the number of genes present on a chip. If some state marker is available to allow the experimenter to determine which genes are changing between particular states, i.e., highly metatstatic versus nonmetastatic disease, then the number of genes that might be designated as interesting targets for prediction could be whittled down to an even smaller number. Still, evaluations with hundreds of possible predictors and tens of targets represent a serious computational challenge. Feasibility at this scale has been tested. Using a multiple processor cluster and computer code optimized for parallel execution, it is possible to run the set of single, double, and triple gene predictions for 60 genes using a set of 500 predictors in 31 experiments within a week (E. Suh and S. Kim, unpublished observation). Such a rate of analysis makes the undertaking possible and will allow studies of the kinds of patterns frequently seen, to develop heuristics, which could further reduce the number of genes that would be likely to provide a strong prediction for any given gene. The extent of the output from such an evaluation, 400 GB, also requires the development of tools to allow an investigator to filter and search the output in efficient ways and to visualize the sets of genes that show codetermination in ways that will help identify trends among the connections.

Finally, we recognize that the use of the proposed analysis is totally independent of the mechanism accounting for the predictive power of a set of genes or conditions. The mechanism producing the association is not a factor, only the ability to predict the expression level of a target

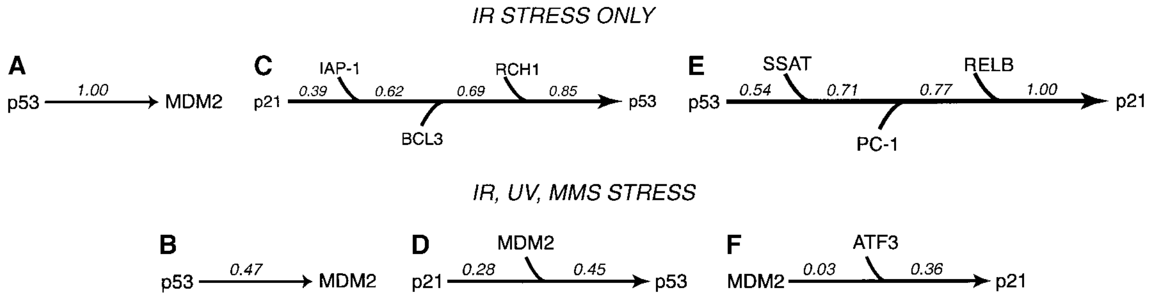


FIG. 7. Determination diagrams for MDM2, p53, and p21 when a single stress or multiple stesses are considered.

gene from the predictor gene levels. The reach of this form of analysis is essentially as broad as the network of interactions in a cell. Genes whose transcription levels are set by closely coupled activities will obviously be predictive of one another's state, whether they are upstream or downstream in the network. Predictions may be possible with codetermined genes from both upstream and downstream of the region of a network acting on the particular target of prediction. In some cases, the predictive genes may be distributed about the genetic network in such a way that their relation to the target gene is based on chains of interaction of various intermediate genes. Thus, whatever the relationship of the predicting genes to the predicted, if knowledge of their states allows us to predict the expression level of the target gene better, then we infer that there is some codeterminative relationship—the better the prediction, the stronger the relation. These key strengths, the ability to discern and rank connections independent of a model of interaction or complete information, are fully aligned with our need to elucidate control and function relationships of newly discovered genes before the complete catalogue of genes is available.

ACKNOWLEDGMENTS

The authors wish to thank Drs. Albert Fornace Jr. and Sally Amundson of the National Cancer Institute for insightful suggestions on the areas of cellular response to DNA damage where perceptron analysis could be fruitful. We are also indebted to Drs. Edward Suh and Robert Martino of the Center for Information Technology of the NIH for access to the Biowolf computer cluster and for parallelization of code.

REFERENCES

Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.* 17–28.

Amundson, S. A., Bittner, M., Chen, Y., Trent, J., Meltzer, P., and Fornace, A. J., Jr. (1999). Fluorescent cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses. *Oncogene* **18**(24): 3666–3672.

Astola, J. T., and Kuosmanen, P. (1999). Representation and optimization of stack filters. *In* "Nonlinear Filters for Image Processing" (E. R. Dougherty and J. T. Astola, Eds.), SPIE Press and IEEE Press, Bellingham, WA.

Bishop, C. M. (1995). "Neural Networks for Pattern Recognition," Oxford Clarendon, New York, and Oxford Univ. Press, London.

Chen, B. P., Liang, G., Whelan, J., and Hai, T. (1994). ATF3 and ATF3 delta Zip. Transcriptional repression versus activation by alternatively spliced isoforms. *J. Biol. Chem.* **269**(22): 15819–15826.

Chen, B. P., Wolfgang, C. D., and Hai, T. (1996). Analysis of ATF3, a transcription factor induced by physiological stresses and modulated by gadd153/Chop10. *Mol. Cell. Biol.* **16**(3): 1157–1168.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* **2**(4): 364–374.

Cohen, D. R., and Curran, T. (1988). fra-1: A serum-inducible, cellular immediate-early gene that encodes a fos-related antigen. *Mol. Cell. Biol.* **8**(5): 2063–2069.

Creemers, J. W., Roebroek, A. J., and Van de Ven, W. J. (1992). Expression in human lung tumor cells of the proprotein processing enzyme PC1/PC3. Cloning and primary sequence of a 5 kb cDNA. *FEBS Lett.* **300**(1): 82–88.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**(4): 457–460.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338): 680–686.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**(25): 14863–14868.

Evan, G., and Littlewood, T. (1998). A matter of life and cell death. *Science* **281**(5381): 1317–1322.

Gorlich, D., Kostka, S., Kraft, R., Dingwall, C., Laskey, R. A., Hartmann, E., and Prehn, S. (1995). Two different subunits of importin cooperate to recognize nuclear localization signals and bind them to the nuclear envelope. *Curr. Biol.* **5**(4): 383–392.

Gorospe, M., Shack, S., Guyton, K. Z., Samid, D., and Holbrook, N. J. (1996). Up-regulation and functional role of p21Waf1/Cip1 during growth arrest of human breast carcinoma MCF-7 cells by phenylacetate. *Cell Growth Differ.* **7**(12): 1609–1615.

Ivanov, V. N., Deng, G., Podack, E. R., and Malek, T. R. (1995). Pleiotropic effects of Bcl-2 on transcription factors in T cells: Potential role of NF-κ B p50-p50 for the anti-apoptotic function of Bcl-2. *Int. Immunol.* **7**(11): 1709–1720.

Kim, S., Dougherty, E. R., Bittner, M. L., Chen, Y., Sivakumar, K., Meltzer, P., and Trent, J. M. (2000). A general nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *J. Biomed. Optics,* in press.

Liang, S., Fuhrman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 18–29.

Liou, H. C., Sha, W. C., Scott, M. L., and Baltimore, D. (1994). Sequential induction of NF-κ B/Rel family proteins during B-cell terminal differentiation. *Mol. Cell. Biol.* **14**(8): 5349–5359.

McAdams, H. H., and Shapiro, L. (1995). Circuit simulation of genetic networks. *Science* **269**(5224): 650–656.

Rosenblatt, F. (1962). "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms," Spartan Books, Washington, DC.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235): 467–470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**(20): 10614–10619.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**(12): 3273–3297.

Szallasi, Z. (1999). Genetic network analysis in light of massively parallel biological data acquisition. *Pac. Symp. Biocomput.* 5–16.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* **22**(3): 281–285.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in *Saccharomyces cerevisiae. Nat. Biotechnol.* **15**(13): 1359–1367.

Yuh, C. H., Bolouri, H., and Davidson, E. H. (1998). Genomic *cis*-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science* **279**(5358): 1896–1902.